ZooKeys 1233: 15–30 (2025)
DOI: 10.3897/zookeys.1233.140726

Short Communication

Extracting specimen label data rapidly with a smartphone—a great help for simple digitization in taxonomy and collection management

Dirk Ahrens[1], Alexander Haas[2], Thaynara L. Pacheco[1], Peter Grobe[1]

1 *Museum A. Koenig, Leibniz Institute for the Analysis of Biodiversity Change (LIB), Adenauerallee 127, 53113 Bonn, Germany*

2 *Museum of Nature Hamburg, Leibniz Institute for the Analysis of Biodiversity Change, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany*

Corresponding author: Dirk Ahrens (D.Ahrens@leibniz-lib.de, ahrens.dirk_col@gmx.de)

Abstract

We provide short tutorials in how to read out specimen label data from type- as well as handwritten labels in a rapid and easy way with a mobile phone. We apply them in general, but test them in particular for insect specimen labels, which are generally quite small. We provide alterative procedure instructions for Android and Apple based environments, as well as protocols for single and bulk scans. We expect that this way of data capture will be of great help for a simple digitization in taxonomy and collection management, independent from large industrial digitization pipelines. By omitting the step of taking/maintaining images of the labels, this approach is more rapid, cheaper, and environmentally more sustainable because no storage with carbon footprint is required for label images. We see the biggest advantage of this protocol in the use of readily available commercial devices, which are easy to handle, as they are used on a daily basis and can be replaced at relatively low cost when they come into (informatic) age, which is also a matter of cyber security.

Key words: Artificial intelligence, citizen science, collection digitization, data science, label transcription, labels, taxonomic impediment, taxonomic revisions

Academic editor: Pavel Stoev
Received: 1 November 2024
Accepted: 17 January 2025
Published: 26 March 2025

ZooBank: https://zoobank.org/70B1D529-3A50-414A-8DAE-D76BFE019BC1

Citation: Ahrens D, Haas A, Pacheco TL, Grobe P (2025) Extracting specimen label data rapidly with a smartphone—a great help for simple digitization in taxonomy and collection management. ZooKeys 1233: 15–30. https://doi.org/10.3897/zookeys.1233.140726

Introduction

Currently, there are immense efforts underway to digitize natural history collections on a large scale, including the associated information and metadata (e.g., Smith and Blagoderov 2012; Hardisty et al. 2020a, b; Belot et al. 2023; Groom et al. 2023; Ong et al. 2023). In these endeavors, among other things, the automatic capture of label data plays a central role (e.g., Beaman et al. 2006; Heidorn and Wei 2008; Lafferty and Landrum 2009; Granzow-de la Cerda and Beach 2010; Haston et al. 2012; Agarwal et al. 2018; Alzuru et al. 2019, 2020; Alzuru 2020; Owen et al. 2020; Belot et al. 2023; Zhang 2023; Takano et al. 2024). However, many of these very promising activities have been for a long time exclusive to large companies, museums or institutions with specialized technical infrastructure, and specially trained staff (e.g., Blagoderov et al. 2012) for the highly customized implementations used (e.g., https://picturae.com/).

Most of the current digitization initiatives aim at a one-go retro-digitization of large collections (Engledow et al. 2018; Hardisty et al. 2020a, b; Helminger et al. 2020; De Smedt et al. 2024). However, this approach comes with limitations: 1) collections are continuously growing and developing (see also Balke et al. 2013); and 2) the scientific community produces a large amount of high-quality biodiversity data independently of the collection institutions with their ongoing research on the specimens, in which amateur scientists are also largely involved (Löbl et al. 2023). The latter is achieved by the often-remote study of the collection material, far away from collections and large digitization pipelines. Especially in insects, taxonomic specialists are rare, and specimens are often loaned by shipment overseas to obtain best identifications from world-leading specialists. In this, working processes are quite different from those of vertebrates or plants in often leading in new methodologies, such as large-scale digitization. However, these data often do not end up in big data repositories, also due to the lack of time and stimulus, as well as the work-overload of taxonomists.

Therefore, more flexible solutions are needed that allow for more efficient data processing and speed up biodiversity/species discovery and help to overcome taxonomic impediment. This would be perfectly in line with the idea of integrating specimen databases and revisionary systematics (Schuh 2012). Advantages of a revision-based digitization (see also Meier and Dikow 2004), i.e., that biodiversity data come from taxonomic revisionary studies rather than from uncritical retro-digitizing of museum specimen data, are the following (based on Meier and Dikow (2004) and Schuh (2012)): 1) the data are provided in association with the most accurate identifications, 2) the data have the most complete taxonomic and geographic coverage, 3) the data satisfy these points in a cost-effective way, and 4) the occurrence data and images are citable and acknowledgeable. This last point enables errors to be retracted and corrected).

Recently, we found that mobile devices, which nowadays are used by almost everyone, can be of assistance in speeding up data collection and digitization, including biodiversity discovery. By simple experimentation, we discovered that mobile phones can be used in association with cloud-like environments (such as Google Workspace or Apple iCloud). Because we think that these workflows will be useful to a larger audience, we prepared this short paper on how to rapidly and easily read out specimen label data using a smartphone.

Most digitization approaches capture digital metadata (e.g., labels) with the intermediate step of digital images (Nelson et al. 2012). This comes with difficulties and the considerable cost of image processing and storage (Tann and Flemons 2008; Hardisty et al. 2020b). It is more sustainable to skip this step, with the greatest benefit to cost, if data can be read and spell-checked at the same moment without the burden of intermediate images, which are scientifically and practically unnecessary for non-type specimens.

Material and methods

Resources needed

1) A mobile or smartphone; a recent model with macrophotography capacity.
2) A stable internet connection.

3) A computer connected with internet and logged into a Google account (via Google Chrome Browser) or AppleID account.
4) A database or text file to insert the specimen data.
5) "*Google Lens*" or "*Google Translate*" to be installed on the mobile phone.

For our testing here, we used a Motorola Edge 30 Neo (system: Android ver. 14), a Motorola G5g Plus (system: Android ver. 10/11), a Motorola G82 G5g (system: Android ver. 13), and an iPhone 15 Pro Max (system: iOS ver. 17.7).

We explored data extraction from labels with different approaches and alternative label conditions (Fig. 1). Some of the tutorials for each of these approaches can be found to be more suitable for different technical situations of the user than others. We describe subsequentially each of these approaches in simplified step by step tutorials. Tutorials are accompanied by screen shots and examples of retrieved data.

With any computer operational system and an Android mobile device:

Variant 1

1) Open the "*Google Translate*" or "*Google Lens*" apps in your mobile phone.
2) Focus on the label to be scanned and zoom virtually by using the touch screen so that the label fills the screen as much as possible. There is no need to be perfectly focused, but all letters should be recognizable.
3) Scan by clicking on the circle with the magnifier lens on it.
4) Mark the label text (Fig. 1) via cursor selection using the touch screen of the mobile phone.
5) Select “Copy to Computer”.
6) Confirm the selected device (computer with which you are logged into your Google account and via Google Chrome Browser) by choosing “Select”.[1]
7) On the computer: simply paste from clipboard into your target document (verbatim label citation).
8) Finally, you may proofread the scan (while having still your specimen in front of you) and manually correct misspellings or readings.
9) Finished.

Alternatively, in step 5 you may choose “Copy” and then paste this copied content into an open Google Docs document on the mobile device. This document can be accessed on via the same Google account on the synchronized computer. This step is sometimes necessary if the internet connection is too slow (see results below). This also works outside of the Google Cloud environment but is a little more complex: Files can be shared between Android, Windows or Mac devices using the "*KDE Connect*" app (https://kdeconnect.kde.org). The latter app works also on Linux. All devices must be in the same WIFI network. After installing the "*KDE Connect*" app, the text can be transferred to the computer.

[1] This function works exclusively with the Google Chrome web browser.

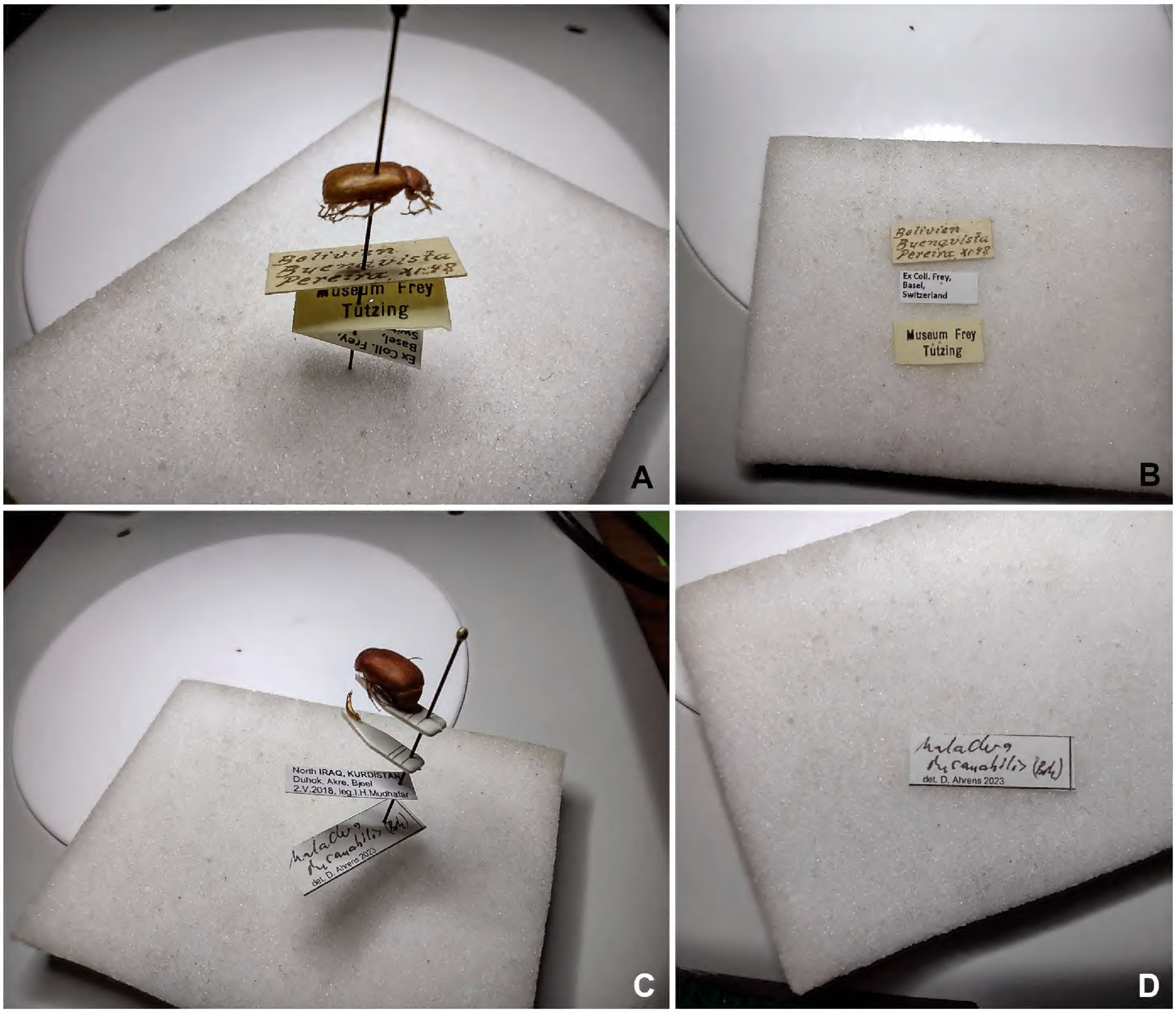

Bolivien Buenavista Pereira XI.48
Museum Frey Tutzing
Ex Coll. Frey, Basel, Switzerland
North IRAQ, KURDISTAN Duhok, Akre, Bjeel 2.V.2018, leg.I.H.Mudhafar
det. D. Ahrens 2023
A
B
C
D

Figure 1. Exemplary specimens used for experimental real-time label scans **A** printed labels scanned on pin **B** printed labels scanned separately **C** partly handwritten labels scanned on pin **D** partly handwritten labels scanned separately.

Variant 2 (bulk scans)

1) Open the "*Google Keep – Notes and Lists*" app on your mobile phone.
2) Click on “+” and then on “image” icon.
3) Click on “Take photo” to capture the image, focus on the label you want to scan. Click on “photo” button, and then on the “checkmark” icon to save it.
4) Click on the image, then on the three dots in the upper right corner, and then on “grab image text”. The text will appear as a note and can be manually corrected for spellings or readings errors. A title for the note can be added. This function seems to work only on newer Android systems; here we used successfully Android ver. 13 and ver. 14. With an older Android ver. 10 or 11 smartphone, this option did not work.
5) Repeat steps 2–4 for each label you want to scan. They will be saved as separate notes.
6) Select all notes, click on the three dots in the upper right corner, and then on “copy to Google Docs” (This step can be alternatively done already

on the computer via the respective google account; see Fig. 5). A single Google Docs document containing all images and texts will be generated. This step can be done on your mobile or on a computer logged into your Google account.

7) On your computer: open your Google Docs file, and the final corrections can be made and downloaded.

With an “Apple-only” environment

Requirements: Make sure you have a recent iPhone or iPad model with macro photography capabilities and the most recent operating system (iOS 15 and later). You will also need a Mac computer and an Apple iCloud account (at least the free version). An internet connection of the phone (e.g., via WLAN) is not necessary for data collection, if you collect your data from the specimen labels first on your phone (bulk scans) and go back to your Mac computer later.

a) Using "*Notes*" app:

1) Open the "*Notes*" app on your iPhone and set up a new note for your current project.
2) In your note, tap the camera symbol at the bottom and choose “scan text” from the pop-up menu. A camera window opens in the bottom part of your note.
3) Aim your camera at the text block you want to scan. Yellow brackets will show you which text block the software sees as target. Once the desired target text is within the brackets press the insert button at the bottom of the camera window. The targeted text will be read and automatically transferred to your note.
4) Briefly check the result in your note.
5) Go to the next line in your note and scan the next target text in the same way, thus accumulating information from multiple specimen labels or multiple specimens as you like.
6) Once finished with the data collecting, return to your desktop Mac computer. If the phone had telephone connection with your provider while you took the scans or on your way back to your desktop computer, the "*Notes*" app should automatically synchronize with your Apple Account in the background so that when you open the "*Notes*" app on your desktop computer, you should find all the scanned data there.
7) Continue to copy and paste the information accumulated in your "*Notes*" app to the document or database of your choice.

b) Using the Shortcuts app:

The "*Shortcuts*" app of iOS can be used to program an automated process from taking the photo, extracting the text and filling a table in Apple’s spreadsheet app "*Numbers*". Make sure that your "*Shortcuts*" and "*Numbers*" apps are synchronized for all of your devices via your iCloud drive. We assembled a "*Shortcuts*" algorithm as a proof of concept. Fig. 4 shows the algorithm.

Without internet connection, using Bluetooth with a Windows PC and an Android 14 mobile phone

1) Download the app ("*Google keeps – Notes and Lists*") on the mobile phone
2) Open Bluetooth options in the computer
3) Pair the computer and mobile phone
4) Click on receive files via Bluetooth
5) Open the app and click on the picture icon
6) Click on "take photo" and take the photo
7) Click on the captured picture
8) Click on the three dots in the upper right corner
9) Click on "grab image text" and select the extracted text[2]
10) Click on the three dots in the lower right corner and click on "send"
11) Click on "send via other apps" and choose the Bluetooth symbol
12) Choose a folder to save the html file in the computer
13) Copy the text from the html-file into a text editor for final spelling corrections

We expect this approach to work in a similar way also in Linux and Apple environments.

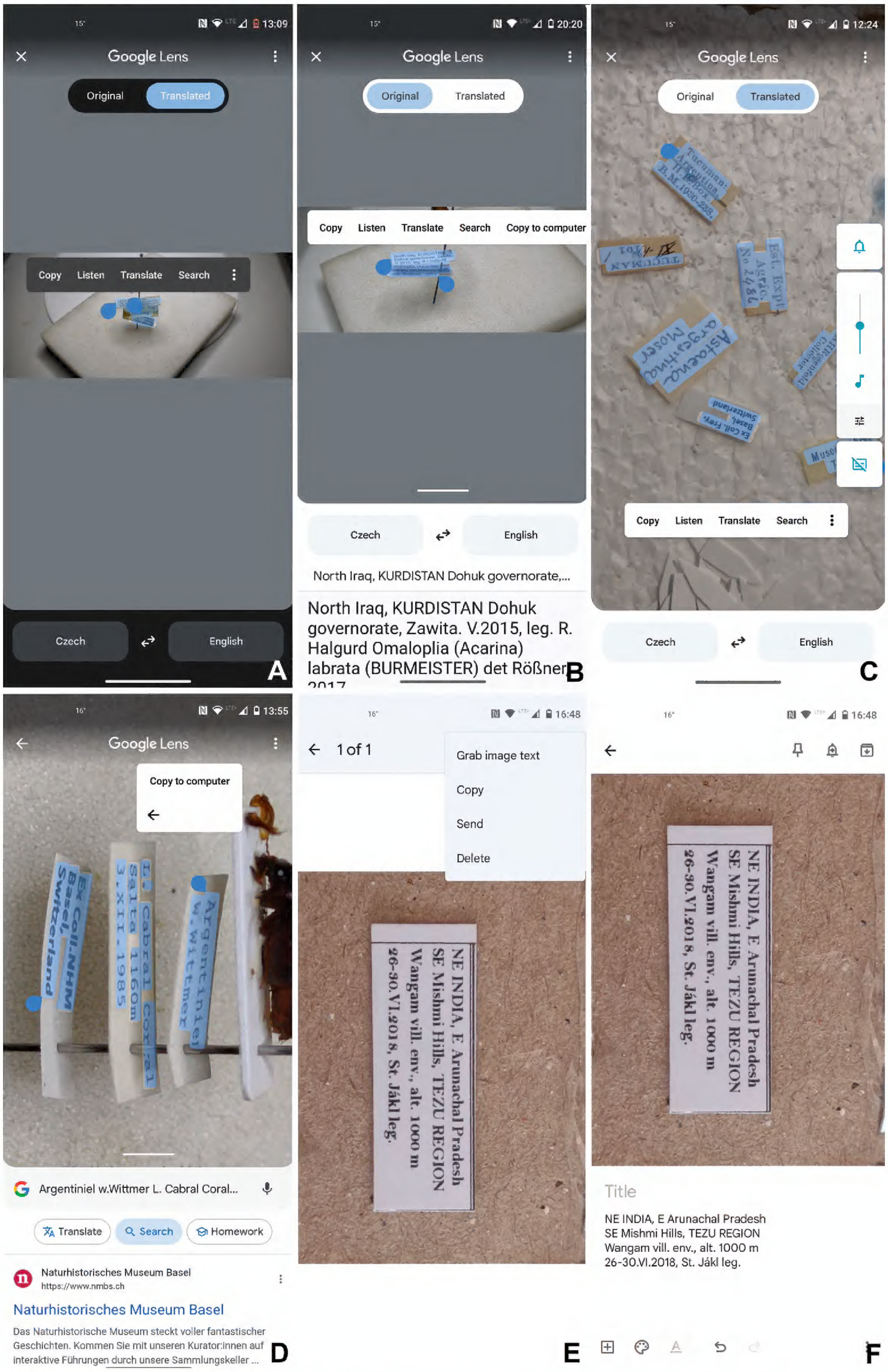
Results

In Table 1 we summarize the major characteristics of data capture using these methods. We show directly pasted content and the amount of real-time spelling corrections needed. In type-printed labels the corrections were minimal, but handwritten labels needed more correcting, depending on the size and style of the handwriting. In these cases, scanning the labels separately from the pin without distortion helped a lot (Fig. 1A, B). In printed labels, direction (Fig. 2C) and distortion of labels did not matter much (Fig. 2D). We were able to scan up to three labels (from the distorted side view) still mounted on a pin and without rotating the labels on, or removing them from, the pin (Fig. 2D).

Low image resolution was not a problem, and we could zoom-in digitally, so the labels almost filled the screen of the phone. However, during our initial testing, we found that much smaller images were also successful in capturing data (Fig. 1A–D).

Processing time per specimen was fast, and we estimated that full data capture, including spelling corrections, was 3–10 seconds per specimen. Processing time was often a little longer for badly handwritten labels, when an insect pin or other labels covered parts of the label text, or when the internet connection was slow. The total time gain per label was larger with labels containing much information or with multiple labels. For example, in the labels shown in Fig. 2B, typing the data by hand into the computer required 60 seconds, including spell check, but scanning the label using Approach 1 took 8 seconds (including a manual spell check). For the data of Fig. 2C, manual typing and spell-check required 121 seconds, while Approach 1 took 10 seconds (including manual spell check).

[2] This function works only on newer Android systems; here we used successfully Android v. 13 and v. 14. With an older Android v. 10 or 11 smartphone, this option unfortunately did not work.



North Iraq, KURDISTAN Dohuk governorate,...

North Iraq, KURDISTAN Dohuk governorate, Zawita. V.2015, leg. R. Halgurd Omaloplia (Acarina) labrata (BURMEISTER) det Rößner

Argentiniel w.Wittmer L. Cabral Coral...

Naturhistorisches Museum Basel
https://www.nmbs.ch

Naturhistorisches Museum Basel

Das Naturhistorische Museum steckt voller fantastischer Geschichten. Kommen Sie mit unseren Kurator:innen auf interaktive Führungen durch unsere Sammlungskeller ...

NE INDIA, E Arunachal Pradesh
SE Mishmi Hills, TEZU REGION
Wangam vill. env., alt. 1000 m
26-30.VI.2018, St. Jákl leg.

Figure 2. Steps of scanning (exemplified by a screenshot from mobile phone) of real-time data collection, and examples of labels **A** step 1: marking of the text to be captured via touch screen of the mobile phone (example - printed labels scanned on pin) **B** step 2: select from menu bar (at the right side under three dots) “Copy to computer” (example - printed labels scanned separately). As to be seen, different labels at different levels on the pin can be scanned simultaneously and do not need to be removed from the pin **C** Screenshot showing the capture of multidirectional printed labels scanned separately from the specimen in "*Google Lens*" **D** Screenshot showing the capture of multiple distorted, printed labels scanned on the pinned specimen in "*Google Lens*" **E** Screenshot showing the initial capture of a printed label scanned separately from the specimen in "*Google Keep*" **F** Screenshot showing the extracted data resulting from **E**.

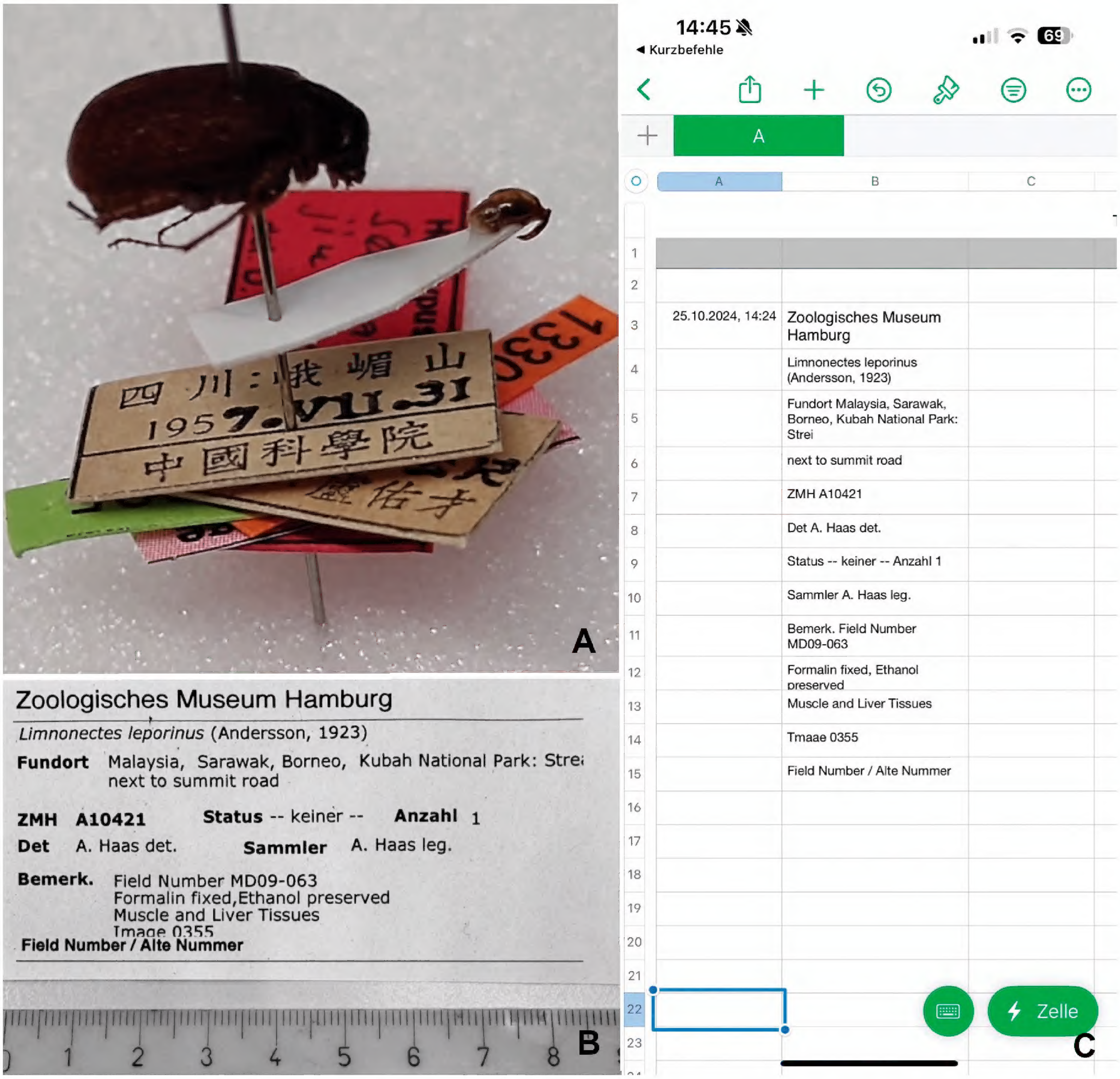
Table 1. Summary of label configuration/ view (with reference to Figs 1, 2) and the obtained resulting text in the final database. Text corrected by real-time manual corrections are indicated in bold.

Label configuration/ view	Text as pasted from computer's clipboard	Verbatim finalized data (after manual correction)
Fig. 1A (labels scanned on pin, distorted)	Belivr vista Peretra、インタ Museum Frey Tutzing Ex Coll. Frey, Basel, Switzer	"B**olivia Buenavista P**ere**i**ra **XI.48** / Museum Frey Tutzing/ Ex Coll. Frey, Basel, Switzer***land***" (CF).
Fig. 1B (labels scanned separately, not distorted)	Bolivia Buengvista Pereira X198 Ex Coll. Frey, Basel, Switzerland Museum Frey Tutzing	"Bolivia Buen**a**vista Pereira X**I.4**8 / Ex Coll. Frey, Basel, Switzerland/ Museum Frey Tutzing" (CF).
Fig. 1C (partly handwritten labels scanned on pin, distorted)	North IRAQ, KURDISTAN Duhok, Akre, Bjeel 2.V.2018, leg.1.H.Mudhafar Maladera del. D. Ahrens 2023	"North IRAQ, KURDISTAN Duhok, Akre, Bjeel 2.V.2018, leg.1.H.Mudhafar/ Maladera **insanbilis (Brsk.)** de**t**. D. Ahrens 2023"
Fig. 1D (partly handwritten labels scanned separately, not distorted)	Maladus dusanabilis (Boy) det. D. Ahrens 2023	Malad**era** **in**sanabilis (B**rsk**) det. D. Ahrens 2023
Fig. 2C	Tucuman: Argentina. H.E.Box. B.M.1930-238. Est. Expt Agric. No 2486 TUCUMAN 101/ AHRosenfeld Collector Astaena argentina Moser Ex Coll. Frey, Basel, Switzerland Museum Frey Tutzing	"Tucuman: Argentina. H.E.Box. B.M.1930-238./ Est. Expt**.** Agric. No 2486/ TUCUMAN **XI-I** 1**9**1/ A H Rosenfeld Collector/ Astaena argentina Moser/ Ex Coll. Frey, Basel, Switzerland/ Museum Frey Tutzing"
Fig. 2D	Argentiniel w.Wittmer L. Cabral Coral Salta 1160 m 3.XII.1985 Ex Coll.NHM Basel, Switzerland	"Argentinie**n** **W.** Wittmer/ L. Cabral Coral Salta 1160 m 3.XII.1985/ Ex Coll. NHM Basel, Switzerland" (NHMB)
Fig. 3A	四川：峨嵋山ج~ 19573131 中國科學院	"四川：峨嵋山 1957.**VII.**31 中國科學院"

We refrained from more extensive comparisons in the time needed between manual typing and data scanning, because the former depends much on the typing skills of the person doing the data entry. The comparative numbers from a few test replicates given here, refer to a scientist without proficiency in typing (performed by D.A.).

In some instances, in Approach 1, we had to use the deviation via a Google Docs document due to bad internet connection, when the copy process failed due to slow data transfer. This was then usually two "clicks" (or seconds) slower, but not a major delay compared to the amount of time required for manual typing.

The iPhone workflow test with was done with a larger label (Fig. 3B). In the workflow 2a) above, using the Notes app on the Apple iPhone, the image recognition tried to identify and focus blocks of text within the label, but did not to capture the label as a whole. To capture multiple bits of information the process had to be reiterated accordingly. Once the data has been collected in

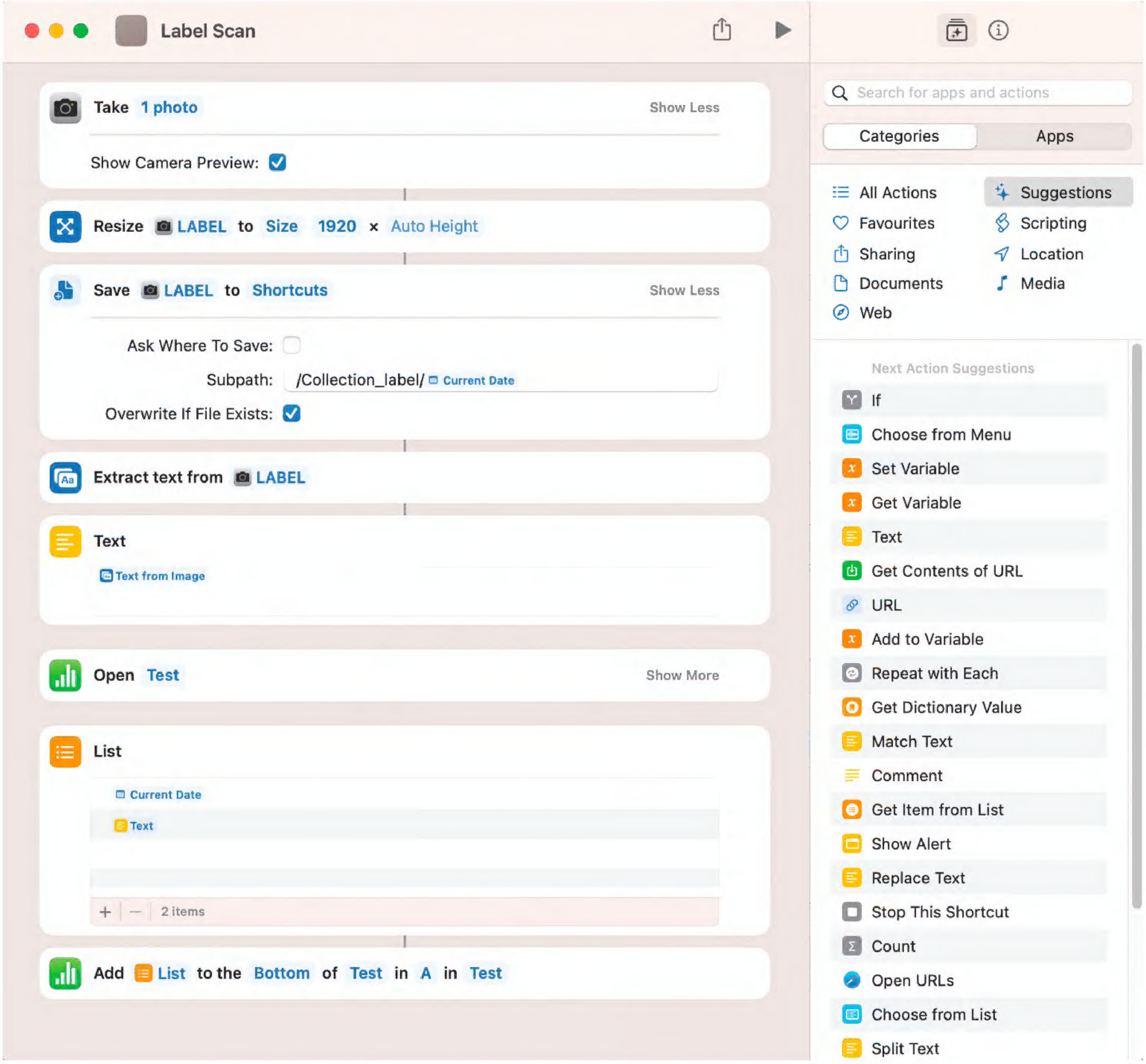


Zoologisches Museum Hamburg
Limnonectes leporinus (Andersson, 1923)
Fundort Malaysia, Sarawak, Borneo, Kubah National Park: Strei
next to summit road
ZMH **A10421** **Status** -- keiner -- **Anzahl** 1
Det A. Haas det. **Sammler** A. Haas leg.
Bemerk. Field Number MD09-063
Formalin fixed,Ethanol preserved
Muscle and Liver Tissues
Image 0355
Field Number / Alte Nummer

	A	B	C
1			
2			
3	25.10.2024, 14:24	Zoologisches Museum Hamburg	
4		Limnonectes leporinus (Andersson, 1923)	
5		Fundort Malaysia, Sarawak, Borneo, Kubah National Park: Strei	
6		next to summit road	
7		ZMH A10421	
8		Det A. Haas det.	
9		Status -- keiner -- Anzahl 1	
10		Sammler A. Haas leg.	
11		Bemerk. Field Number MD09-063	
12		Formalin fixed, Ethanol preserved	
13		Muscle and Liver Tissues	
14		Tmaae 0355	
15		Field Number / Alte Nummer	

Figure 3. Other exemplary specimens used for experimental label scans **A** for Chinese language labels (printed) **B** printed Herpetology collection label that was scanned in the test of the Apple "*Shortcuts*" app algorithm. Note the incomplete text in the third text line and the cut off text "image 0355" below (compare to the corresponding data entries in C) **C** Screenshot of the automatically scanned collection label as transferred into cells of the spreadsheet app "*Numbers*". Although the text scan was very reliable, incomplete text will need editing: the somewhat cut off text "image 0355" of the label was interpreted as "Tmaee 0355". The time stamp in the first column corresponds to the file name of the respective photo saved as backup in the Shortcuts directory.

Notes, further copy-paste editing is necessary to transfer the data to a database. Workflow 2b), using "*Shortcuts*" app automation (Fig. 4) scans the whole label and also stores the data with a timestamp directly into a spreadsheet app. Furthermore, the photos are stored in the user's Apple iCloud account (as backups for potential later reference), but this step is optional in the algorithm. The result of the scanning is shown in Fig. 3B, C. Note that incomplete text in the original caused interpretation problems (truncated third line and partially hidden bottom part of "image 0355"). In addition, the algorithm tends to place



Take 1 photo Show Less

Show Camera Preview: ☑

Resize LABEL to Size 1920 × Auto Height

Save LABEL to Shortcuts Show Less

Ask Where To Save: □

Subpath: /Collection_label/ Current Date

Overwrite If File Exists: ☑

Extract text from LABEL

Text

Text from Image

Open Test Show More

List

Current Date

Text

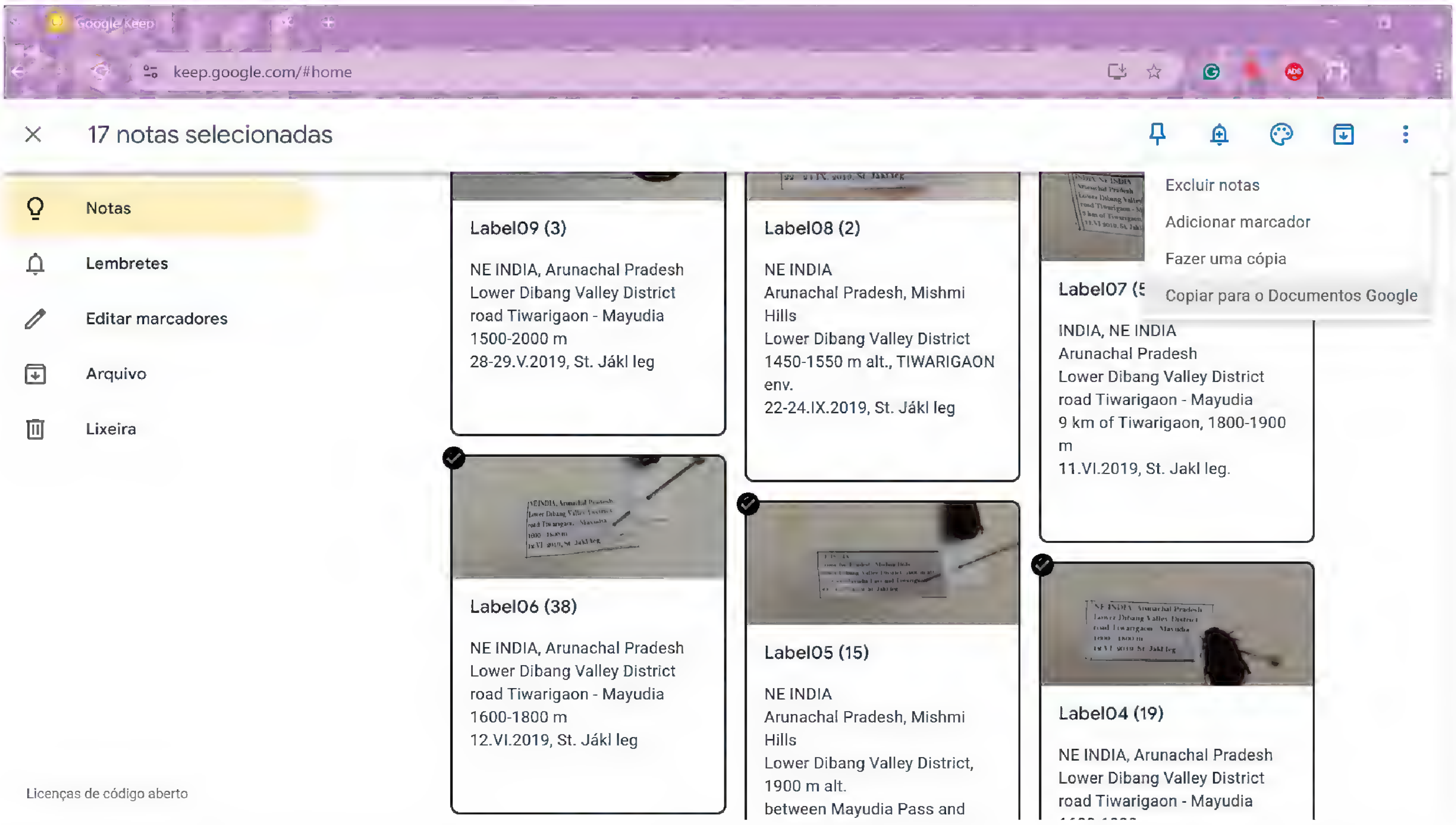
+ — 2 items

Add List to the Bottom of Test in A in Test

Figure 4. iOS "*Shortcuts*" app algorithm. From top to bottom: The first step will open the iPhone's Camera app and lets you photograph the label. The photo ("LABEL") is then resized (optional, to reduce space) and saved in the background to the "*Shortcuts*" directory in your iCloud account with the current date (and time) as file name. Then the text is extracted from the photo and stored to a text container. The next step opens the spreadsheet "Test" in app "*Numbers*"; an empty target spreadsheet file (here: "Test") must be prepared beforehand and waiting in the "*Shortcuts*" folder of your iCloud account. Current Date and Text items are then collected in the "List". The List items are finally entered into different columns in the spreadsheet file "Test" and a sheet with the name "A".

each recognized line of text in a separate cell. If several lines belong to the same block of information, editing of the cells was necessary. The scan of the label and filling of the cells in the spreadsheet took less than 10 seconds. The algorithm analyzed the label as lines of text and allocated one cell per line in the spreadsheet. This means that the locality information in our example was split up into two cells in our test. Depending on which further tasks the user wants to accomplish copy-paste processing of such splits is necessary.

The approach using a Bluetooth connection between the mobile phone and the computer appeared to be slightly longer (by the increased number of "device



Label09 (3)

NE INDIA, Arunachal Pradesh
Lower Dibang Valley District
road Tiwarigaon - Mayudia
1500-2000 m
28-29.V.2019, St. Jákl leg

Label06 (38)

NE INDIA, Arunachal Pradesh
Lower Dibang Valley District
road Tiwarigaon - Mayudia
1600-1800 m
12.VI.2019, St. Jákl leg

Label08 (2)

NE INDIA
Arunachal Pradesh, Mishmi Hills
Lower Dibang Valley District
1450-1550 m alt., TIWARIGAON env.
22-24.IX.2019, St. Jákl leg

Label05 (15)

NE INDIA
Arunachal Pradesh, Mishmi Hills
Lower Dibang Valley District,
1900 m alt.
between Mayudia Pass and

Label07 (5

INDIA, NE INDIA
Arunachal Pradesh
Lower Dibang Valley District
road Tiwarigaon - Mayudia
9 km of Tiwarigaon, 1800-1900 m
11.VI.2019, St. Jakl leg.

Label04 (19)

NE INDIA, Arunachal Pradesh
Lower Dibang Valley District
road Tiwarigaon - Mayudia

Figure 5. Screenshot of bulk-scanned labels via "*Google Keep*", inspected afterwards directly from the computer interface, during the step of copying to of the label text to a Google document (interface here in Portuguese).

clicks") than the direct approach (with the internet connection). Yet it saved a substantial amount of time for extracting the label data compared with manually typing. The use of Bluetooth may be necessary in situations where a good internet connection is unavailable, such as in collections. However, unfortunately the "grab image text" function of "*Google Keeps*" did work only with a newer smartphone (Android 13 or 14), not on an older device with Android 10 or 11.

Bulk approaches are available under the Google and Apple environment (Fig. 4) with the "*Google Keeps*" and "*Notes*" applications, respectively. In both, images are temporarily stored in the mobile devices, which can be later saved or discarded. While bulk approaches save time with data transfer, they have the disadvantage that potentially incomplete scans are only discovered when the specimens no longer at hand.

Discussion

While new technologies, including artificial intelligence, are entering in our daily life, their use and application in biodiversity research is yet rather limited, although there have been developments in AI-powered label recognition (Johaadien and Torma 2023; Weaver et al. 2023; Takano et al. 2024). Herein, we addressed the scanning of data from labels using a smartphone using different operating systems. To our knowledge, this has not previously been explored and applied to insect collection specimens in particular. There are solutions for large-scale mass digitization of collections (Blagoderov et al. 2012; Tegelberg et al. 2014; Engledow et al. 2018; Belot et al. 2023). All these solutions require manual separation of specimens and labels to separately photograph them. Initial trials with robotic technology (e.g. Dupont and Price 2019) are promising, but such methods can only be used by larger institutions with the budget to do so.

With partly omitting the so far obligatory step of taking and permanently storing images of the labels, our direct approach to data capture is more rapid and environmentally more sustainable. In some of our procedures, data extraction happens without delay in the background, and there still is the option to retain the images if wanted. For a simple extraction of distributional data for taxonomic revisions or faunistic studies, we see no scientific necessity for long-term storage of images of specimen labels. Moreover, spell checking of the scanned and extracted data can be done when the specimen is at hand, and the data is finalized almost immediately.

However, depending on the individual needs and working conditions, the user has the choice on the individual workflow. It is possible to scan 50 labels in a row (i.e., bulk workflow) before transferring the data to the computer. Then in some critical cases, having a backup photo is good for quality assessment and spell checking.

Another great advantage is that these protocols use commercial devices that are simple to handle and which cost little to replace when they come into (informatic) age which is also a matter of cybersecurity. Unfortunately, in biosystematics, specialized devices are often overpriced, technologically obsolete, or require often expensive updates and service. Since biodiversity research in invertebrates, and especially entomology, is done in part by amateur scientists (and even professionals may lack funds for their "descriptive research"), funding may be lacking or limited.

Our results revealed that some functions may not be available on older smartphones running earlier operating systems. Here, the "grab image text" function did not work on Android 11 or older. Similarly, the other approaches might not work with even older devices, even if they work with the same app. However, we could not explore these limitations in detail, because we had only a limited selection of smartphones at hand.

The increasingly high reliability of text recognition and the rapid data transfer may make the use of machine readable barcode labels and QR codes superfluous in collection management, since connected data can be easily inferred from numerical voucher numbers on labels. Considering that optical character recognition (OCR) software, even when coupled with very advanced AI-technologies, possibly might give more errors than reading machine-readable codes (barcodes, 2D-codes), more rigorous tests are needed to check and compare the accuracy of a smartphone-based workflow compared to standard barcode and QR readers in this field.

Our solutions and tutorial proposed here are well suited for fast, secure recording of collection objects, e.g. when visiting a collection or when selecting individual objects. We are aware that habits, skills, and specific workflows influence the way we integrate such devices and text recognition capabilities. We are convinced that they will make a significant contribution and help to alleviate the taxonomic impediment (e.g., de Carvalho et al. 2005, 2007; Engel et al. 2021), as the workload for taxonomists capturing data of the material they study will be reduced by at least tenfold.

Finally, we note that there might be even more options and possibilities to scan labels with mobile devices. These options might evolve as quickly as mobile phones and artificial intelligence technology improve. Nevertheless, we expect that our paper will be an inspiration to others to continue exploring options on how to successfully apply this technology in their workflows and to share what they have learnt.

Acknowledgements

We thank the numerous colleagues who provided insightful discussions and encouraged us to pursue this topic. Furthermore, we thank Ilia Gjonov, Salza Palpurina, and Pavel Stoev for their helpful comments on the former version of our manuscript, as well as our colleague Christina Blume (LIB Bonn) for helping with some of the test runs of the apps.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Funding

No funding was reported.

Author contributions

All authors have contributed equally.

Author ORCIDs

Dirk Ahrens https://orcid.org/0000-0003-3524-7153
Alexander Haas https://orcid.org/0000-0002-3961-518X
Thaynara L. Pacheco https://orcid.org/0000-0001-9503-7751
Peter Grobe https://orcid.org/0000-0003-4991-5781

Data availability

All of the data that support the findings of this study are available in the main text.

References

Agarwal N, Ferrier N, Hereld M (2018) Towards automated transcription of label text from pinned insect collections. IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV (USA), 2018. Institute of Electrical and Electronics Engineers, 189–198. https://doi.org/10.1109/WACV.2018.00027

Alzuru I (2020) Human-machine extraction of Information from biological collections. PhD thesis, University Florida, 160 pp.

Alzuru I, Malladi A, Matsunaga A, Tsugawa M, José FAB (2019) Human-Machine Information Extraction Simulator for Biological Collections. IEEE International Conference on Big Data (Big Data), Los Angeles, CA (USA), 2019. Institute of Electrical and Electronics Engineers, 4565–4572. https://doi.org/10.1109/BigData47090.2019.9005601

Alzuru I, Matsunaga A, Tsugawa M, Fortes JAB (2020) General Self-aware Information Extraction from Labels of Biological Collections. IEEE International Conference on Big Data (Big Data), Atlanta, GA (USA), 2020. Institute of Electrical and Electronics Engineers, 3035–3044. https://doi.org/10.1109/BigData50022.2020.9377737

Balke M, Schmidt S, Hausmann A, Toussaint EFA, Bergsten J, Buffington M, Häuser CL, Kroupa A, Hagedorn G, Riedel A, Polaszek A, Ubaidillah R, Krogmann L, Zwick A, Fikáček M, Hájek J, Michat MC, Dietrich C, La Salle J, Mantle B, Ng PKL, Hobern D (2013) Biodiversity into your hands - A call for a virtual global natural history 'metacollection'. Frontiers in Zoology 10: 55. https://doi.org/10.1186/1742-9994-10-55

Beaman RS, Cellinese N, Heidorn PB, Guo Y, Green AM, Thiers B (2006) HERBIS: Integrating digital imaging and label data capture for herbaria [Abstract]. Botany 2006, California State University – Chico. 28 July–2 August 2006. http://www.2006.botanyconference.org/engine/search/index.php?func=detail&aid=402

Belot M, Preuss L, Tuberosa J, Claessen M, Svezhentseva O, Schuster F, Bölling C, Léger T (2023) High Throughput information extraction of printed specimen labels from large-scale digitization of entomological collections using a semi-automated pipeline. Biodiversity Information Science and Standards 7: e112466. https://doi.org/10.3897/biss.7.112466

Blagoderov V, Kitching I, Livermore L, Simonsen T, Smith VS (2012) No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133–146. https://doi.org/10.3897/zookeys.209.3178

de Carvalho MR, Bockmann FA, Amorim DS, de Vivo M, de Toledo-Piza M, Menezes NA, de Figueiredo JL, McEachran JD (2005) Revisiting the taxonomic impediment. Science 307: 353–353. https://doi.org/10.1126/science.307.5708.353b

de Carvalho MR, Bockmann FA, Amorim DS, Brandão CRF, Vivo M de, Figueiredo JL de, Britski HA, Pinna MCC de, Menezes NA, Marques FPL, Papavero N, Cancello EM, Crisci JV, McEachran JD, Schelly RC, Lundberg JG, Gill AC, Britz R, Wheeler QD, Stiassny MLJ, Parenti LR, Page LM, Wheeler WC, Faivovich J, Vari RP, Grande L, Humphries CJ, DeSalle R, Ebach MC, Nelson GJ (2007) Taxonomic impediment or impediment to taxonomy? A commentary on systematics and the cybertaxonomic-automation paradigm. Evolutionary Biology 34: 140–143. https://doi.org/10.1007/s11692-007-9011-6

De Smedt S, Bogaerts A, De Meeter N, Dillen M, Engledow H, Van Wambeke P, Leliaert F, Groom Q (2024) Ten lessons learned from the mass digitisation of a herbarium collection. PhytoKeys 244: 23–37. https://doi.org/10.3897/phytokeys.244.120112

Dupont S, Price BW (2019) ALICE, MALICE and VILE: high throughput insect specimen digitisation using angled imaging techniques. Biodiversity Information Science and Standards 3: e37141. https://doi.org/10.3897/biss.3.37141

Engel MS, Ceríaco LMP, Daniel GM, Dellapé PM, Löbl I, Marinov M, Reis RE, Young MT, Dubois A, Agarwal I, Lehmann PA, Alvarado M, Alvarez N, Andreone F, Araujo-Vieira K, Ascher JS, Baêta D, Baldo D, Bandeira SA, Barden P, Barrasso DA, Bendifallah L, Bockmann FA, Böhme W, Borkent A, Brandão CRF, Busack SD, Bybee SM, Channing A, Chatzimanolis S, Christenhusz MJM, Crisci JV, D'elía G, Da Costa LM, Davis SR, De Lucena CAS, Deuve T, Fernandes Elizalde S, Faivovich J, Farooq H, Ferguson AW, Gippoliti S, Gonçalves FMP, Gonzalez VH, Greenbaum E, Hinojosa-Díaz IA, Ineich I, Jiang J, Kahono S, Kury AB, Lucinda PHF, Lynch JD, Malécot V, Marques MP, Marris JWM, Mckellar RC, Mendes LF, Nihei SS, Nishikawa K, Ohler A, Orrico VGD, Ota H, Paiva J, Parrinha D, Pauwels OSG, Pereyra MO, Pestana LB, Pinheiro PDP, Prendini L, Prokop J, Rasmussen C, Rödel M-O, Rodrigues MT, Rodríguez SM, Salatnaya H, Sampaio Í, Sánchez-García A, Shebl MA, Santos BS, Solórzano-Kraemer MM, Sousa ACA, Stoev P, Teta P, Trape J-F, Van-Dúnem Dos Santos C, Vasudevan K, Vink CJ, Vogel G, Wagner P, Wappler T, Ware JL, Wedmann S, Zacharie CK (2021) The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. Zoological Journal of the Linnean Society 193(2): 381–387. https://doi.org/10.1093/zoolinnean/zlab072

Engledow H, De Smedt S, Groom Q, Bogaerts A, Stoffelen P, Sosef M, Van Wambeke P (2018) Managing a mass digitization project at Meise Botanic Garden: From start to finish. Biodiversity Information Science and Standards 2: e25912. https://doi.org/10.3897/biss.2.25912

Granzow-de la Cerda Í, Beach JH (2010) Semi-automated workflows for acquiring specimen data from label images in herbarium collections. Taxon 59: 1830–1842. https://doi.org/10.1002/tax.596014

Groom Q, Dillen M, Addink W, Ariño AHH, Bölling C, Bonnet P, Cecchi L, Ellwood ER, Figueira R, Gagnier P-Y, Grace OM, Güntsch A, Hardy H, Huybrechts P, Hyam R, Joly AAJ, Kommineni VK, Larridon I, Livermore L, Lopes RJ, Meeus S, Miller JA, Milleville K, Panda R, Pignal M, Poelen J, Ristevski B, Robertson T, Rufino AC, Santos J, Schermer M, Scott B, Seltmann KC, Teixeira H, Trekels M, Gaikwad J (2023) Envisaging a global infrastructure to exploit the potential of digitised collections. Biodiversity Data Journal 11: e109439. https://doi.org/10.3897/BDJ.11.e109439

Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul DL, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020a) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6: e54280. https://doi.org/10.3897/rio.6.e54280

Hardisty A, Livermore L, Walton S, Woodburn M, Hardy H (2020b) Costbook of the digitisation infrastructure of DiSSCo. Research Ideas and Outcomes 6: e58915. https://doi.org/10.3897/rio.6.e58915

Haston E, Cubey RWN, Pullan M, Atkins H, Harris D (2012) Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. ZooKeys 209: 93–102. https://doi.org/10.3897/zookeys.209.3121

Heidorn PB, Wei Q (2008) Automatic metadata extraction from museum specimen labels. In: Greenberg J, Klas W (Eds) Metadata for semantic and social applications: Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, 22–26 September 2008, DC 2008: Berlin, Germany. Universitätsverlag Göttingen, Göttingen, 57–68.

Helminger T, Weber O, Braun P (2020) Digitisation of the LUX herbarium collection of the National Museum of Natural History Luxembourg. Bulletin de la Société des Naturalists Luxembourgeois 122: 147–152.

Johaadien R, Torma M (2023) “Publish First”: a rapid, GPT-4 based digitisation system for small institutes with minimal resources. Biodiversity Information Science and Standards 7: e112428. https://doi.org/10.3897/biss.7.112428

Lafferty D, Landrum LR (2009) SALIX, a semi-automatic label information extraction system using OCR [Abstract]. Botany & Mycology 2009, Snowbird, Utah, 25–29 July 2009. http://2009.botanyconference.org/engine/search/index.php?func=detail&aid=130 [accessed 21.X.2024]

Löbl I, Klausnitzer B, Hartmann M, Krell F-T (2023) The silent extinction of species and taxonomists—an appeal to science policymakers and legislators. Diversity 15(10): 1053. https://doi.org/10.3390/d15101053

Meier R, Dikow T (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. Conservation Biology 18: 478–488. https://doi.org/10.1111/j.1523-1739.2004.00233.x

Nelson G, Paul D, Riccardi G, Mast A (2012) Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19–45. https://doi.org/10.3897/zookeys.209.3135

Ong S-Q, Mat Jalaluddin NS, Yong KT, Ong SP, Lim KF, Azhar S (2023) Digitization of natural history collections: a guideline and nationwide capacity building workshop in Malaysia. Ecology and Evolution 13: e10212. https://doi.org/10.1002/ece3.10212

Owen D, Groom Q, Hardisty A, Leegwater T, Livermore L, van Walsum M, Wijkamp N, Spasić I (2020) Towards a scientific workflow featuring Natural Language Processing for the digitisation of natural history collections. Research Ideas and Outcomes 6: e58030. https://doi.org/10.3897/rio.6.e58030

Schuh R (2012) Integrating specimen databases and revisionary systematics. ZooKeys 209: 255–267. https://doi.org/10.3897/zookeys.209.3288

Smith V, Blagoderov V (2012) Bringing collections out of the dark. ZooKeys 209: 1–6. https://doi.org/10.3897/zookeys.209.3699

Takano A, Cole TCH, Konagai H (2024) A novel automated label data extraction and data base generation system from herbarium specimen images using OCR and NER. Scientific Reports 14(1): 112. https://doi.org/10.1038/s41598-023-50179-0

Tann J, Flemons P (2008) Data capture of specimen labels using volunteers. Australian Museum. http://australianmuseum.net.au/Uploads/Documents/23183/Data%20Capture%20of%20specimen%20labels%20using%20volunteers%20-%20Tann%20and%20Flemons%202008.pdf [accessed 21.X.2024]

Tegelberg R, Mononen T, Saarenmaa H (2014) High-Performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. Taxon 63(6): 1307–1313. https://doi.org/10.12705/636.13

Weaver WN, Ruhfel BR, Lough KJ, Smith SA (2023) Herbarium specimen label transcription reimagined with large language models: capabilities, productivity, and risks. American Journal of Botany 110(12): e16256. https://doi.org/10.1002/ajb2.16256

Zhang Y (2023) Use of artificial intelligence (AI) in historical records transcription: Opportunities, challenges, and future directions. Master thesis, McGill University, 24 pp.